Curso fundamenteal de Inferencia Filogenética Molecular

Pablo Vinuesa (vinuesa@ccg.unam.mx)

Progama de Ingeniería Genómica, CCG, UNAM

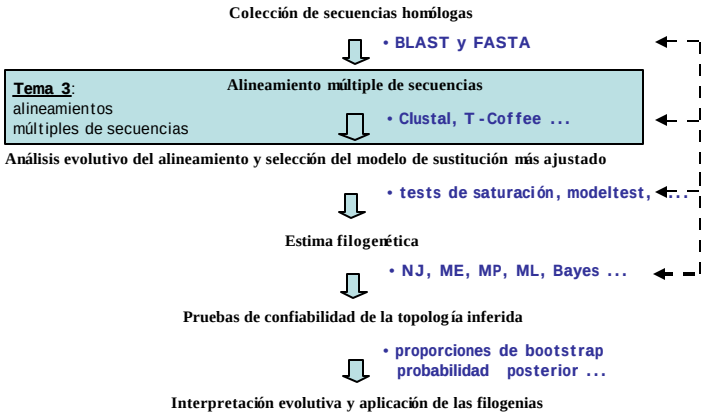
http://www.ccg.unam.mx/~vinuesa/

Tutor: PDCBM, Ciencias Biológicas, PDCBioq. y
Profesor de la Lic. Ciencias Genómicas y posgrado

- **Tema 3: Alineamientos múltiples**

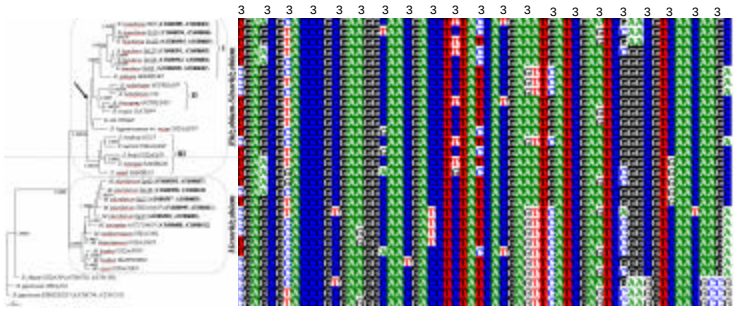
1. Alineamientos múltiples y el problema de las repeticiones, sustituciones e indeles
2. Alineamientos múltiples progresivos usando programas de la familia Clustal
3. Formatos de secuencia
4. Alineamiento de secuencias codificadoras de proteínas usando RevTrans y DAMBE
5. Alineamiento de genes ribosomales usando RDP-II y GreenGenes

Protocolo básico para un análisis filogenético de secuencias moleculares



Tema III: alineamientos múltiples

- Cualquier estudio de filogenético o de evolución molecular basado en secuencias necesita de un **alineamiento múltiple para determinar las correspondencias de homología a nivel de los resíduos individuales o caracteres**.
- La mejor manera de representar un alineamiento múltiple es escribiendo las secuencias a comparar en filas una encima de la otra, generándose una matriz de m x n (secs. x posic) caracteres, en la que cada columna contiene a resíduos homólogos

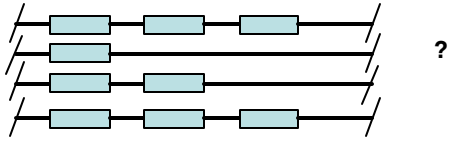


- Comparar los aln. múltiples en el contexto de una filogenia nos puede revelar mucho acerca de los patrones y tasas de sustitución.

Tema III: alineamientos múltiples –

- **El problema de las repeticiones**

Muchas **proteínas multidominio** pueden presentar diverso grado de **repetición de dominios** particulares. Puede llegar a ser muy complejo o prácticamente imposible hacer el alineamiento correcto de estos "repeats".



A nivel de DNA se dan también regiones repetidas, muchas veces involucrando a unos pocos nts. como es el caso de los **microsatélites y otras regiones repetidas.** Con frecuencia estas regiones son imposibles de alinear objetivamente. Suelen acumularse en regiones no codificantes del genoma, o en regiones codificantes hipervariables como espaciadores intergénicos transcritos o regiones reguladoras (UTRs).

Este tipo de "repeats" cortos son poco frecuentes a nivel de aminoácidos, si bien a este nivel es común encontrar regiones o dominios "de gran escala" repetidos. Un ejemplo clásico de este fenómeno son las calmodulinas.

Tema III: alineamientos múltiples –

• El problema de las sustituciones

• Al examinar alns. múltiples de proteínas se obaservan dos patrones de sustitución:

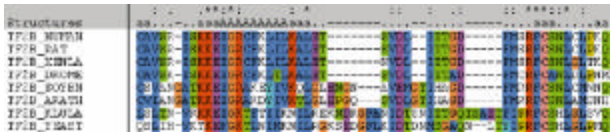
1.- Existen bloques de 5 a 20 resíduos con alto nivel de identitad y similitud dispersos entre regiones de menor similitud. Estos bloques corresponden típicamente a elementos estructurales como **α-hélices y pliegues beta** que evolucionan más lentamente que los loops o bucles que los interconectan

2.- Las columnas alineadas con múltiples estados de caracter tienden a presentar resíduos de con características bioquímicas similares (I, A, V, L; S, T; R, K; etc.). Esta conservación de resíduos similares es particularmente patente en los bloques correspondientes a elementos de estructura secundaria, sitios activos o de unión a ligandos.
La propiedad bioquímica más conservada es la de polaridad/hidrofobicidad.

Tema III: alineamientos múltiples –

• El problema de las sustituciones

• Es importante recordar que por debajo del **20% de identidad** a nivel de sec. de AA es ya imposible que se pueda obtener un alineamiento múltiple (o pareado) confiable si nos basamos para obtenerlo sólo en la secuencia primaria, puesto que entramos en la **zona de penumbra**



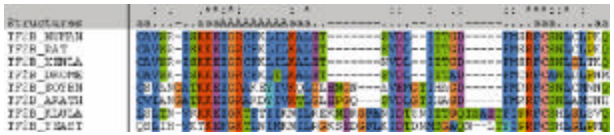
• Un par de secuencias de nts al azar presentarán en promedio un 25 % de dentidad.

• Por tanto, siempre que sea posible, hay que realizar los alineamientos múltiples en base a las secuencias traducidas, es decir, sobre AAs (igual que al hacer búsquedas en bases de datos de secuencia)

Tema III: alineamientos múltiples –

• El problema de los indeles (inserciones/deleciones)

• Cuando por eventos de inserción o deleción (**indeles**) las secuencias homólogas presentan distintas longitudes, es necesario introducir "**gaps**" en el alineamiento para mantener la correspondencia entre sitios homólogos situados antes y después de las regiones afectadas por indeles. Estas regiones se identifican mediante guiones (-).

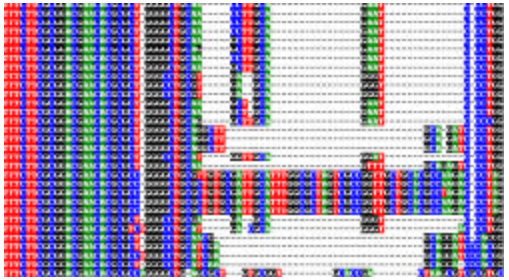


Los indeles no se distribuyen aleatoriamente en las secuencias codificadoras.
Casi siempre aparecen ubicados entre dominios funcionales o estructurales, preferentemente en bucles (loops) que conectan a dichos dominios.
Esto vale tanto para RNAs estructurales (tRNAs y rRNAs) como para proteínas.
No suelen interrumpir el marco de lectura.

• Generalmente se usan **sistemas de penalización de gaps afines** (GP = gap + (ext. x long.))

Tema III: alineamientos múltiples –

• A mayor **distancia genética** (evolutiva) entre un par de secuencias, mayor será el número de mutaciones acumuladas. Dependiendo del tiempo de separación de los linajes y la tasa evolutiva del locus, puede llegar a ser imposible alinear ciertas regiones debido a fenómenos de **saturación mutacional**. En loci de evolución muy rápida como intrones o espaciadores intergénicos, los fenómenos de saturación mutacional se observan incluso cuando se comparan secuencias de organismos evolutivamente próximos (mismo género o familia).



¡Las regiones de homología dudosa deben de ser excluídas de un análisis filogenético!
Debemos de procurar maximizar la relación entre señal/ruido

Alineamientos múltiples (AM)

- Existen diversos algoritmos (además de matrices de sustitución y esquemas de "gap penalty") para la generación de AMs. Unos son **exahaustivos** (garantizan encontrar el alineamiento óptimo) y otros son **heurísticos** (no lo garantizan)

- No existe un algoritmo ideal para todas las situaciones. Para búsquedas en bases de datos se emplean algoritmos heurísticos para encontrar **alineamientos locales** (FastA y BLAST). Para análisis filogenéticos necesitamos métodos que produzcan **alineamientos globales**.

- Algoritmos basados en **programación dinámica** (PD) aseguran encontrar la solución óptima o el mejor **alineamiento global** para 2 secuencias. Se trata de un algoritmo $O(N^2)$, ya que el tiempo y memoria que demandan es proporcional al producto de las long. de ambas secuencias (N1 X N2). Se puede generalizar el proceso para la comparación de múltiples secuencias, usando la **función de objetividad** llamada **suma ponderada de pares (WSP):**

$$SS\ W_{ij}\,D_{ij}$$

Donde D_{ij} es la puntuación de cada posible par de secuencias y W_{ij} es un factor de ponderación arbitrario que permite dar más o menos peso a ciertas comparaciones (por ej. en función de su score D_{ij}. Algoritmos de PD se pueden emplear para encontrar el AM que da el mejor valor posible de la función WSP. El problema radica en que la complejidad crece exponencialmente con cada nueva secuencia que se añade (complejidad O(NM)), donde N=long. sec M= no. secs. Ello implica que se alcanza rápidamente un límite computacional

Alineamientos múltiples (AM)

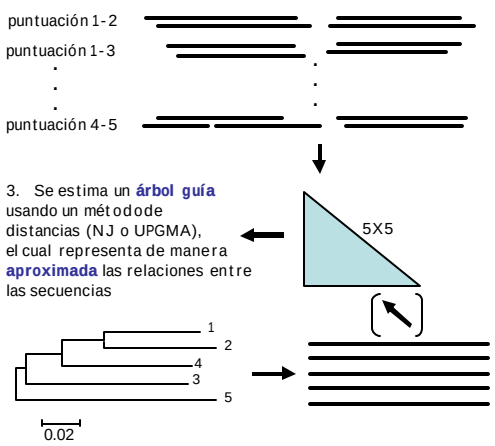
Existen diversas estrategias computacionales para obtener alineamientos múltiples de manera (semi)automática.

1.- Implementación de **algoritmos de alineamiento progresivo**.

Así como los alns. múltiples son indispensables para reconstruir filogenias a partir de secs, un árbol de relaciones filogenéticas representa información muy valiosa para guiar la generación de un aln. múltiple.

La mayor parte de los alineadores automáticos modernos se basan en este tipo de algoritmos. Construyen un árbol guía aproximado a partir de distancias calculadas entre todos los pares posibles de secuencias. De la matriz de distancias resultantes se construye un árbol usando un método algorítmico (NJ o UPGMA). El árbol guía resultante se emplea para construir el alineamiento de manera progresiva. Las dos secuencias más similares se alinean primero usando DP y una matriz o esquema de ponderación particular. Una vez alineado el primer par, los gaps generados ya no se mueven. Este par es tratado como una sola secuencia y es alineada contra la siguiente secuencia o grupo de secuencias más próximas en el árbol. Se repite el proceso hasta que todas las secs. están alineadas. El proceso es suficientemente rápido como para alinear varios cientos de secuencias. Son menos precisos que los métodos basados en la WSPs, pero muchísimo más rápidos.

Pasos en la generación de un alineamiento múltiple siguiendo la estrategia de alineamiento progresivo



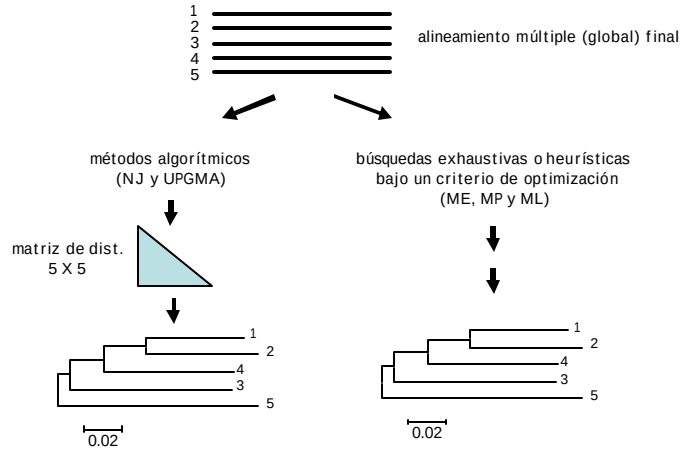
1. Se generan todos los posibles **alineamientos pareados**, usando métodos heurísticos o exhausivos (**PD**), y se calcula su score (puntuación) en base a la matriz de sustitución y gap penalties elegida

2. Se calcula una **matriz de distancias** en base a las puntuaciones de los alineamientos pareados del paso anterior

3. Se estima un **árbol guía** usando un método de distancias (NJ o UPGMA), el cual representa de manera **aproximada** las relaciones entre las secuencias

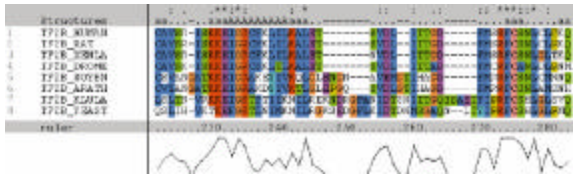
4. Se hace el **alineamiento** riguroso (PD) y **global entre pares de secuencias** siguiendo el orden de similitud indicado por el árbol guía

Pasos en la generación de un alineamiento múltiple siguiendo la estrategia de alineamiento progresivo - y su uso para estimar una filogenia



Alineamientos múltiples progresivos usando Clustal

- **La familia** Clustal es posiblemente la más popular para hacer AMs de nt y aa
- Existen versiones para todas las plataformas y en red (http://www.ebi.ac.uk.clustalw)
- La primera versión (Clustal) salió en 1988, la última, **ClustalX**, en 1997 (última Vers. = 1.83)
- **ClustalX (X-windows Clustal)** lee secuencias en diversos formatos, calcula un **árbol guía NJ**, usando algoritmos heurísticos o exhaustivos sobre aln. locales basado en **distintas matrices de pesado y de penalización de gaps afines y sitio-específicos**. Puede hacer **alineamientos de perfiles** y existen diversas **herramientas de control de calidad** del AM. Permite incluir criterios estructurales para guiar el AM, usando **máscaras estructurales**. Partes del alineamiento o secuencias particulares pueden ser **realineadas** para ir obteniendo un aln global cada vez mejor. Es decir, ClustalX no sólo genera alineamientos (como ClustalW), sino que éstos pueden ser editados y mejorados interactivamente por el usuario. Además, ClustalX (y ClustalW) permite la **reconstrucción y visualización de árboles NJ** y hacer **análisis de bootstrap** sobre los alineamientos. Finalmente, los AMs pueden ser escritos en **diversos formatos de salida** (CLUSTAL, FASTA, NEXUS, PHYLIP ...)

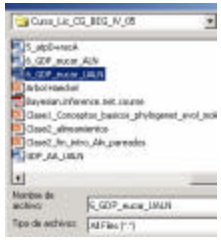


Alineamientos múltiples progresivos usando Clustal -aspectos prácticos

- Para obtener un AM con clulstal tenemos que tener todas las secuencias homólogas en un solo archivo. Estas secs. pueden estar escritas en diversos formatos (FASTA, EMBL SWISS-PROT ...)
- Sobre este archivo se puede correr un primer análisis usando las opciones por defecto de Clustal
- Según el grado de divergencia de las secuencias a analizar, puede ser muy útil probar distintas series de matrices y valores de gap penalty. Existen scripts de Perl que prueban sistemáticamente una gran cantidad de combinaciones de parámetros para encontrar aquellos que maximizan el score del alinemiento (**MULTICLUSTAL**). Yuan et al., 1999 BioInformatics 15:862-863.
- Clustal es adecuado para alinear sets de secuencias totalmente colineares (no usar para ensamblar contigs!) y que presentan el mismo órden de dominios estructurales
- **Condiciones en las que Clustal no puede operar de manera óptima**

1. Si tenemos unas pocas secuencias muy divergentes de una superfamilia; ajustar "delay parámeter" y/o usar modo de alineamiento de perfiles, preferentemente con máscara estructural
2. Sesgo composicional en aas hidrofílicos (G, P, S, N, D, Q, E, K, R) pueden introducir demasiados gaps (penalidades de indel sitio-específico)

Alineamientos múltiples progresivos usando Clustal -un ejemplo: alineamiento de GDPs dependientes de NAD



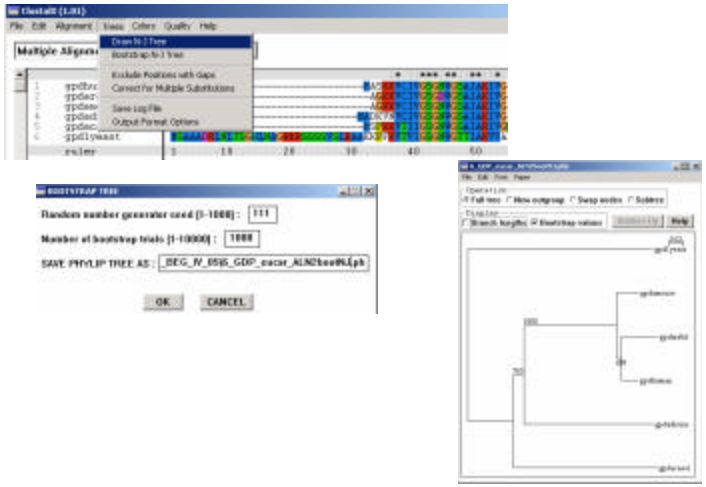
1.- Seleccional modo de aln y fichero a alinear (en este caso las secs. están escritas en formato FASTA)

```
>gpdlyeast
MSAAADRLNLTSGHLNAGRKRSSSSVSLKAAEKPFKVTVIGSGNWGTTIA
KVVAENCKGYPEVFAPIVQMWVFEEEINGEKLTEIINTRHQNVKYLPGIT
LPDNLVANPDLIDSVKDVDIIVFNIPHQFLPRICSQLKGHVDSHVRAISC
LKGFEVGAKGVQLLSSYITEELGIQCGALSGANIATEVAQEHWSETTVAY
HIPKDFRGEGKDVDHKVLKALFHRPYFHVSVIEDVAGISICGALKNVVAL
GCGFVEGLGWGNNASAAIQRVGLGEIIRFGQMFFPESREETYYQESAGVA
DLITTCAGGRNVKVARLMATSGKDAWECEKELLNGQSAQGLITCKEVHEW
LETCGSVEDFPLFEAVYQIVYNNYPMKNLPDMIEELDLHED
>gpdadrome
MADKVNVCIVGSGNWGSAIAKIVGANAAALPEFEERVTMFVYEELIDGKK
LTEIINETHENVKYLKGHKLPPNVAVPDLVEAAKNADILIFVVPHQFIPN
FCKQLLGKIKPNAIAISLIKGFDKAEGGGIDLISHIITRIPCAVLMGANL
ANEVAEGNFCETTIGCTDKKYGKVLRDLFQANHFRVVVVDDADAVEVCGA
LKNIVACGAGFVDGLKLGDNTKAAVIRLGLMEMIRFVDVFYPGSKLSTFF
ESCGVADLITTYRRVSEAFVTSGKTIEELEKEMLNGQKLQGPPTAEEVNY
```

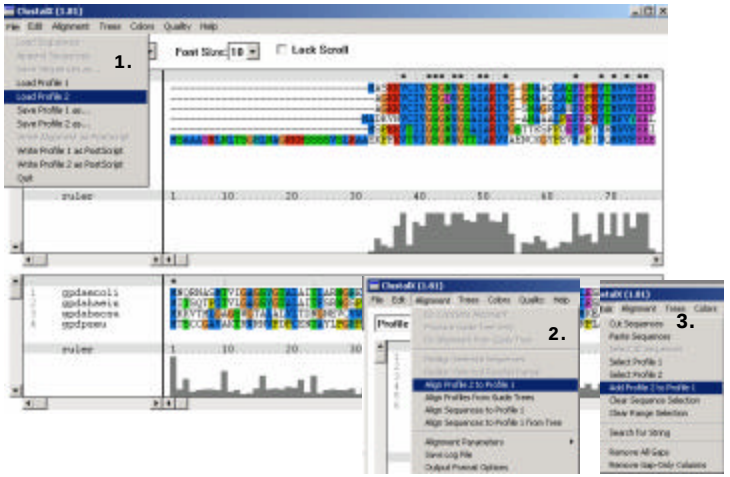
...

Alineamientos múltiples progresivos usando Clustal -un ejemplo: alineamiento de GDPs dependientes de NAD

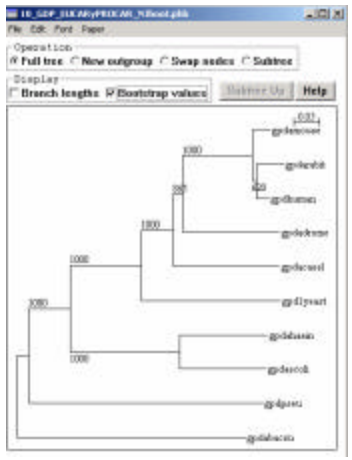
Alineamientos múltiples progresivos usando Clustal -un ejemplo: alineamiento de GDPs dependientes de NAD



Alineamientos múltiples progresivos usando Clustal -un ejemplo: alineamiento de GDPs dependientes de NAD



Alineamientos múltiples progresivos usando Clustal -reconstrucción de una filogenia (NJ) mediante NJplot



Servidores para alinear nts. en base a un alineamiento de proteínas

¡¡¡ **Siempre que quieras alinear secs. de DNA codificadoras (CDSs) alinea primero sus productos y usa el alineamiento múltiple de proteínas para guiar el de los genes correspondientes** !!! Usa para ello servidores como protal2dna o RevTrans, o tus propios scripts de Perl

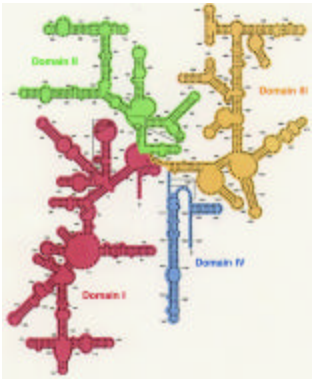
http://bioweb.pasteur.fr/seganal/interfaces/protal2dna.html

Servidores para alinear nts. en base a un alineamiento de proteínas

http://www.cbs.dtu.dk/services/RevTrans/

Servidores para alinear secuencias de rRNAs o rDNAs

- Los genes ribosomales representan un problema muy particular en el contexto de alineamientos múltiples. Deben de guiarse usando máscaras de información estructural.



- Servidores como **GreenGenes** y **RDP-II** proveen herramientas muy útiles en este contexto. Si quieres ver unos tutoriales sobre el uso de estos servidores, visita mi sito web y busca bajo phylogeny tutorials:
  http://www.ccg.unam.mx/~vinuesa/Using_the_GreenGenes_and_RDPII_servers.html

Formatos de secuencias
I) FASTA

- Existen una gran cantidad de estilos o formatos de presentación de secuencias. Muchos programas de análisis filogenético usan su propio formato (Phylip, Nexus, Mega ...)

- El formato más sencillo es el **FASTA**, en el que cada secuencia se identifica mediante un renglón descriptor que comienza con **>** en el siguiente renglón comienza la secuencia

```
>R._galegae
CCGCTGGTCACCTCCGGCAAGCGCGCCATCCACCAGGAAGCGCCTTCCTA
CGTCGATCAGTCGACCGAAGGCCAGATCCTGGTCACCGGCATCAAGGTCG

>M._plurifarium
CCGGTCGACGCCGTCGAGCTGCGTGCCATCCACCAGCCGGCTCCGGCCTA
TGTCGACCAGTCGACGGAAGCGCAGATCCTGGTTACCGGCATCAAGGTTC

>B._japonicum
CCGGTCAAGTCGGAAGGCCTGCGCGCCATCCACCAGGAAGCGCCGACCTA
CACCGACCAGTCCACCGAAGCTGAAATTCTCGTCACCGGCATCAAGGTCG
```

Formatos de secuencias
II) PHYLIP

- **Phylip (interleaved):** no. seqs, no. caracteres
  nombre secuencias (máx 10 caracteres) espacio, secuencia ...

```
 3      100
R._galegae CCGCUGGUCA CCUCCGGCAA GCGCGCCAUC CACCAGGAAG CGCCUUCCUA
M._plurifa ...G.C.A.G ..GU..AGCU ...U...... ......CCG. .U..GG....
B._japonic ...G.CAAGU .GGAA...CU .......... .......... ....GA....

           CGUCGAUCAG UCGACCGAAG GCCAGAUCCU GGUCACCGGC AUCAAGGUCG
           U.....C... .....G.... CG........ ...U...... ........UC
           .AC...C... ..C....... CUG.A..U.. C......... ..........
```

- **Phylip (sequential or non-interleaved)**

```
 3 100
R._galegae   CCGCTGGTCA CCTCCGGCAA GCGCGCCATC CACCAGGAAG CGCCTTCCTA
             CGTCGATCAG TCGACCGAAG GCCAGATCCT GGTCACCGGC ATCAAGGTCG
M._plurifa   CCGGTCGACG CCGTCGAGCT GCGTGCCATC CACCAGCCGG CTCCGGCCTA
             TGTCGACCAG TCGACGGAAG CGCAGATCCT GGTTACCGGC ATCAAGGTTC
B._japonic   CCGGTCAAGT CGGAAGGCCT GCGCGCCATC CACCAGGAAG CGCCGACCTA
             CACCGACCAG TCCACCGAAG CTGAAATTCT CGTCACCGGC ATCAAGGTCG
```

Formatos de secuencias
III) NEXUS

```
#NEXUS
[OJO!!!, no usar guiones -, sólo guiones bajos_]

BEGIN TAXA;               [taxa block]
DIMENSIONS NTAX=3;
TAXLABELS
R._galegae;
M._plurifarium;
B._japonicum;
END;

BEGIN CHARACTERS;         [character block]
DIMENSIONS NCHAR=100;
FORMAT DATATYPE=DNA  MISSING=? GAP=- MATCHCHAR=. INTERLEAVE=yes ;
MATRIX
[                         10        20        30        40        50]
[                          *         *         *         *         *]
R._galegae      CCGCTGGTCACCTCCGGCAAGCGCGCCATCCACCAGGAAGCGCCTTCCTA
M._plurifarium  ...G.C.A.G..GT..AGCT...T............CCG..T..GG....
B._japonicum    ...G.CAAGT.GGAA...CT.......................GA....

[                         60        70        80        90       100]
[                          *         *         *         *         *]
R._galegae      CGTCGATCAGTCGACCGAAGGCCAGATCCTGGTCACCGGCATCAAGGTCG
M._plurifarium  T.....C........G....CG...........T.............TC
B._japonicum    .AC...C.....C.......CTG.A..T..C..................
;
END;
```

Formatos de secuencias: su interconversión

- Cuando preparamos un fichero con nuestras propias secuencias generalmente lo más adecuado es hacerlo en formato FASTA
- Si necesitamos pasarlo a otro formato, una buena posibilidad es hacerlo con **ReadSeq**

  http://iubio.bio.indiana.edu/cgi-bin/readseq.cgi

ReadSeq reconoce automáticamente el formato de entrada y si se trata de aas o nts

- Muchos de los paquetes de software que utilizaremos en el curso tales como BioEdit, ClustalX, DAMBE, MEGA3 y PAUP* son capaces de leer e interconvertir diversos formatos